

AD-401445

(SP Series)



SP-1083

The Microstatistics of Text

Lauren B. Doyle

February 21, 1963

SYSTEM DEVELOPMENT CORPORATION, SANTA MONICA, CALIFORNIA

February 21, 1963

1
(page 2 blank)

SP-1083

THE MICROSTATISTICS OF TEXT

ABSTRACT

This paper is a reappraisal of the role of statistics in text analysis. Current inhibiting influences in the use of statistics are discussed. The question of descriptive vs. predictive statistics is explored at some length. A distinction between macrostatistics and microstatistics is made, with the implication that the former should be used in describing libraries whereas the latter should be used in describing written language.

The second section of the paper pictures a relationship between the probability of occurrence of a word or word group in text and the cognitive effect of such a word or word group. This relationship is then illustrated through statistical data on word pairs; statistics of pairs which are directly linked in a sentence-structure tree are compared to statistics of pairs which, though the words are adjacent in text, are not directly linked in such a tree. This study of statistics as a function of sentence structure is then extended to units of text larger than a word pair.

The final section discusses the problem of selecting and displaying content-indicative word groups in condensed representations of documents. It explains why the statistical approach, by itself or in conjunction with other techniques, is unavoidable in a problem such as automatic abstracting, and illustrates the perils faced by some non-statistical methods which have been talked about in the recent literature.

I. The Meaning of Statistical Treatment

More than five years have passed since H. P. Luhn's "A Statistical Approach to Mechanized Encoding and Searching of Literary Information" was published (1). One offshoot of the statistical approach--automatic abstracting--has attracted attention and follow-up research, but in general statistical treatment of text has not gathered the momentum it deserves.

There appear to be at least three inhibiting factors:

- 1) Very large samples of text are needed before statistical treatment can begin to look impressive, but even at this late date one does not see many million-word corpora dotting the countryside, nor the means to process them easily.
- 2) Automatic abstracting, the first major point of application of the statistical approach, has not covered itself with glory; unfortunately, many people blame the statistical approach rather than unrealistic expectations. As with any tool, there is a limit to what statistics can do.
- 3) Statistical treatment seems like an indirect and unnatural method of language analysis. Syntactical analysis, on the other hand, seems like a direct method and therefore a natural method. Many linguists hold this view, and their expressions of it are characteristically persuasive.

These three factors, especially the latter, have conspired to scare people away from the statistical path, and as a result statistics is a fringe methodology in language data processing when perhaps it might be a central methodology. To better understand the intellectual climate, we need to take a look at the misgivings expressed by those who feel that only detailed syntactic and semantic analysis can permit meaningful handling of language by automatic processes.

A typical argument goes like this: "How can you satisfactorily cope with natural language by pure, unaided statistical analysis? Processing text is not like dealing with urns filled with various colored balls, nor like analyzing neutron attenuation in a crystal lattice; one is dealing with words--with meaning-bearing elements in context. Words in text provide you with much more information than you would ordinarily get in a situation where statistics is applied, and your statistical processes don't get the benefit of this information."

Unfortunately, such an argument has the psychological effect of rejecting statistics rather than of indicating its proper use, and furthermore the argument is relevant only to some problems in language data processing--not to all.

The most significant thing about statistical techniques is their ease of application. To extract the full information from a text one needs some kind of semantic analysis, and today this must be done by a human on a word-by-word basis; automation of this kind of analysis is a long way off. But automation of statistical analysis is well within the current state of the art of text processing. It is so easy that it makes a great deal of sense to explore statistical techniques just to find out how much they can accomplish.

Furthermore, an easy analysis does not have to be a pure, unaided statistical analysis. Those aspects of syntactic or semantic analysis which are already automatic could be brought in to increase the accuracy of the over-all analysis. For example, if one has developed an automatic method of distinguishing between the article "a" and the mathematical symbol "a," there is no reason why it shouldn't be used to aid statistical analysis of article usage. In any event, the potentialities of currently feasible computer language-handling techniques are so great that we must develop them even in the face of demonstrable crudeness.

are we sampling?

Ironically, the small band of protagonists of the statistical approach is contributing to the inhibited climate, even though unwittingly. In particular, overemphasis on refined statistical methodology has given an awesome look to this kind of language analysis. If this is inhibiting, it is needlessly so, because sensibly applied simple arithmetic can give better results in language analysis than woodenly applied advanced techniques. Sophisticated measures are often applied "because they are there" rather than because they are appropriate, and as we shall see in subsequent paragraphs a questionable light is being shed on a whole class of measures which statisticians instinctively use.

Also, the practitioners of statistical analysis have underplayed the theme of G. K. Zipf (2) that the output of text-analyzing statistical procedures is strongly related to what language is. If such a relationship exists, it guarantees that statistical analysis can give useful information both for theoretical and practical purposes. Indeed, this article is being written partly to illustrate some of the kinds of information about language which can be had, much of it through application of the most elementary of statistical measures, such as totals and percentages. Some of this information may even be startling in its significance for language data processing.

Prior to such illustrations, we need to understand the meaning of statistical treatment of text, and to thereby understand why it is inevitable in the course of development of language data processing. As a corollary to this we need to realize what statistics will probably not do for us.

To delve into the latter point, consider that there are two ways of applying statistics, predictively and descriptively. The predictive function was put by A. F. Parker-Rhodes (3) as follows: "...Statistics is a mathematical technique for the assessment of objectivity in estimates made, on more or less inadequate information, of quantities actually dependent on a supposedly much ampler set of data than is (at) hand...."

Then, however, Parker-Rhodes argues that the predictive function is not necessarily applicable in library classification theory: "...It has been assumed that statistical methods could be applied to the data in such a way as to reveal any objectively existing classes which may be there. The general name for the techniques evolved in this way is factor analysis....I propose to suggest...that some classification problems are in principle outside its scope...." He thereupon implies that the "objectively existing classes" revealed either by a librarian's analysis of a collection of books or by a statistical analysis of text may exist only for the corpus on hand--and hence the statistical analysis cannot be used to predict what other libraries will be like.

This is reasonable because libraries are made up arbitrarily, and we ought not to expect their statistics to reflect the operations of some laws of nature which might, in turn, allow us to predict things about other libraries which we have not analyzed. Parker-Rhodes says, talking about a library category: "...As such, it corresponds to a statistic, an empirically found quantity, rather than to a parameter whose existence is already guaranteed by an underlying theory...."

The assumption made by some in applying statistics to library analysis is that the library is a sample of the universe* of text. But is it? In actuality, we can never know when it is or isn't in any given experiment unless we have deliberately arranged to take a random sample of a universe of text. I have not heard of anyone doing that, or even suggesting doing it; as a matter of fact, since most students of language data processing like to think of universes of text as including some future text as well as past text, such a random sampling would be slightly impossible. Therefore, since we have no way of knowing of what population a real library is a sample, the great bag of tools known as "predictive statistics" does not give us any particular advantage in this situation over other statistics we might use. Of course, one might use correlation coefficients or chi-squared for the sake of convenience--this is tolerable as long as one keeps pinching himself to remind himself that he is not working with a random sample.

Someone might ask, "What's wrong with taking a random sample of a library for making statements about that library?" This is legitimate, but the library

*Or of some part of that universe, e.g., the universe of all biology text.

itself is a biased sample of something bigger, and the question arises as to whether one is any better off with a random sample of a biased library than with a biased sample of the library. Moreover, most workers in the language data processing area seem to prefer to take biased rather than random samples. In machine translation this policy lets one get by with a smaller glossary. And in information retrieval, Swanson (4) argues that if we want to "extrapolate" from a small sample to a large library we would be better off with a highly biased sample than with a random sample. Swanson was concerned with knowing about the fine structure of a library and realized that a random sample wouldn't tell much about that. Note that the "extrapolation" that one hopes to do with biased samples is not the same thing as "prediction." Extrapolation from biased samples has no mathematical basis, nevertheless researchers in this field have no choice but to attempt it.

It is difficult to say what proper role predictive statistics will play in the development of language data processing. Of course, as an instrument in human-factors experiments involving language it will play its usual role. But in analysis of text, where we know little about what influences were dominant in the generation of the text, predictive statistics may be applicable only in a small minority of analyses involving statistical treatment.

As we might expect, many people will use predictive statistics from force of habit and training. But if the showdown comes on the kind of mathematics to use, a blend of descriptive statistics and set theory is a sure-fire bet. Though, as Bar-Hillel (5) emphasizes, the history of application of set mathematics to indexing has been depressing, because of a tendency toward premature theorizing, there is no reason why this should continue to be the case.

descriptive statistics

How do we describe something that is large? One way is to describe it in terms of large elements: in describing a mountain one might say, "It has a crater at the top and is cone-shaped, with a ridge running eastward connecting it to a larger mountain."

Suppose that large elements cannot be perceived. Before weather satellites, the meteorologist was in the position of having to describe something which was so big that he couldn't see its elements. The only procedure open to him was to collect hundreds of small-scale descriptions from various weather stations and piece them together into an over-all description. The large elements, such as cold fronts and air masses, could then be mapped out.

The librarian, like the meteorologist, can't see the large elements of the book collection. From the librarian's viewpoint, the desired large elements are classes of books which have subject matter in common. These classes can be related together to form a classification scheme which, in effect, is a picture or map of a library. But here also, as in the case of the meteorologist,

the whole scheme has to be derived by piecing together hundreds and thousands of small-scale descriptions. Each description is derived from a single book and consists of the decision that this book has enough in common with a group of other books that it belongs on the same shelf with them, i.e., the book should be assigned to a particular category.

Usually, for economic reasons, classification schemes are not derived from each individual library. A library hopes to be able to import a scheme used by other libraries and hopes that this scheme will have categories into which its books will fall with reasonably good distribution. If, as is often the case, the categories of the imported scheme do not become equally stocked, then the scheme is no longer a true picture of the library. However, making the picture true is a simple matter, at least in principle: count the books in each category and make a bar graph, using the classification scheme as the ordinate. This is the most elementary of statistical descriptions, yet it could be very informative to a literature searcher, who could tell at a glance how specialized the library is and what the fields of specialization are.

Ideally we would like a description of a library which is independent of any a priori classification scheme, and it is possible to achieve this through computer-programmed analysis of the raw text of a library. But here also, since a computer cannot "see" the large elements of the library any more than we can, it must go through the procedure of summarizing the myriads of small elements which it can see.

An example of a text-derived statistical description is the well-known rank-frequency curve of G. K. Zipf (2). It would be a simple matter to inventory all of the words in a library and list the words in order of their frequency; computer programs now exist (6) by which this may be done, and which Zipf could well have used had they been available in his times. Such rank-frequency listings, however, would not be tremendously useful as descriptions of libraries, although they are useful as descriptions of individual documents. In fact, the "most frequent word" lists which are used in automatic abstracting and indexing are subsets of a rank-frequency listing.

It is worthy of note that the information of greatest value to us in such rank-frequency descriptions of documents is not the numerical part, but the meanings of the words which turn out to be most frequent. In descriptive statistics, the numeric is often a means to an end, in contrast to predictive statistics, where we are actually interested in estimating some quantity with the aid of statistics. Researchers who form the habit of regarding the numeric as selective rather than informative in itself will be less likely to lose time in pursuit of mathematical rigor and less apt to disdain workable ad hoc formulae which are not found in the tool kit of a professional statistician.

There are many statistical descriptions which are of greater interest to us than rank-frequency lists. In particular, any description which suggests

library structure (as classification schemes do) is potentially useful in information retrieval. In a previous article (7) I have discussed the efforts of myself and others to generate "condensed representations" of libraries directly from text. Automatic generation of such representations is possible as a result of the fact that books on similar topics resemble each other in their word content, as Luhn pointed out (1). The variety of forms which such representations could assume is bewilderingly huge, which of course makes this a fascinating field of exploration; however, as discussed above, the inhibitions on work in this area are great, and the number of workers is still very small, considering both the size of the opportunity and the difficulty of the problem.

microstatistics

In analyzing some of the material presented in Section II of this article, I realized the need to make a distinction between "macrostatistics" and "microstatistics" of text. The distinction revolves around what elements are counted in the process of statistical description. One can count either word tokens or word types.* When one summarizes and lists the most frequent words in a document, he is counting the word tokens--the individual occurrences of words in the text. His summary is made up of word types, e.g.,

fuel	94
ignition	70
kerosene	64
combustion	62
pressure	58
pump	49
volts	47

(etc.)

A word type thus stands for all the replications of itself which are found in the text.

One can count word types also, and when one is describing something as large as a library in terms of its text, it can be expedient to do this. As the first stage in such a description one might want to inventory all the summary lists (like that above) for all of the documents. When one thus "summarizes the summaries," one is enabled to make statements like "out of 10,278 documents, 112 contain the word 'fuel' among their ten most frequent content words, 65 contain the word 'kerosene' among the top ten, and 22 of the latter also contain 'fuel' in the top ten." Summarizing word types in this manner is what I call a macrostatistical procedure, whereas the process of counting the word tokens to form the summary lists of word types I would call a microstatistical procedure.

*For the sake of completeness, the distinction should also take in letter types and letter tokens, since many contemporary statistical analyses of text involve letter counting as well as word counting.

Of course, many statistical researchers are working with index tags assigned to documents by humans, rather than with keyword lists. Is their work macro-statistical? Strictly speaking an index tag is a word token, i.e., it is not an entry on a summary list. However, it is desirable to make sets of index tags and keyword lists conceptually equivalent; both describe a document even though the index tags do not constitute a statistical description. Moreover, both can be utilized in roughly the same way as base material for a statistical description of an entire library. As will become more apparent in subsequent discussion, little is lost by regarding index tags essentially as word types rather than as word tokens.

The requirement for the distinction between "micro" and "macro" becomes a little clearer if we consider an analogy between the basic ideas of mathematical probability and those of descriptive statistics. In probability one chooses to reckon in terms of events which are "equally likely"; for example, in computing dice throw probabilities one starts with the assumption that each face of a dice cube is as likely to turn up as each of the other faces. In descriptive statistics one looks for elements which are of "equal interest." Since equality of interest is determined somewhat subjectively, in contrast to equality of likelihood, one therefore has a little more freedom in deciding what to count in a statistical description; however, such a freedom can be abused, to the detriment of one's description.

The problem of describing a library is a case in point. One would regard the postulate "All documents are created equal" as being a reasonable foundation for a library description. Therefore one would like to count either documents or things which pertain to documents, such as index tags--being careful of course to deal with the same number of index tags for each document.* Obviously, if one decides to describe the library by counting the word tokens of the text as "of equal interest" he will find that documents contribute to the description in proportion to their size, and the postulate "Big documents are more important than little documents" is at odds with "All documents are created equal."

Now, I know of at least one statistical worker who, if he had been asked, would have agreed that "All documents are created equal" is a reasonable postulate. Nevertheless, his description of the library was such that a few documents contributed enormously and many others contributed not at all. (The nil contribution was possible because he chose to count tokens of

*One can use different numbers of index tags per document in a statistical description without violating the equality-of-interest postulate if one assigns a weight to each of a document's tags which is inversely proportional to the number of tags or, in the case of co-occurrence statistics, inversely proportional to one less than the number of tags times the number of tags ($N^2 - N$).

selected word types, and some of his items did not contain any of the selected types.) Could he have recognized the contradiction between "All documents are created equal" and his method? Probably not, because he was using a predictive statistic (the correlation coefficient) and no doubt assumed, as is the mental habit, that he was sampling some universe of text and thereby unearthing parameters. This is a good example of what an intellectual trap a predictive statistical framework can be.

If one is describing a language, rather than a library, one has to be exceedingly careful what one regards as "of equal interest." If the statistical worker above made the mistake of counting tokens when he should have counted types, it is interesting to consider that linguists often make just the opposite mistake in their conception of language: they count types when they should count tokens, which (in view of the fact that linguists are not inclined to count things) translates itself into "they regard types rather than tokens as being of equal interest." This mental habit is what leads to the familiar and frustrating phenomenon of "the counterexample." Though a given counterexample may be quite rare in usage, it can be just as devastating in a linguistic argument as if it were abundant. In fact, I have heard some counterexamples which I am convinced have never been said by anyone aside from the host linguist, but which were nevertheless devastating.

It is asserted that, to the extent statistical description is possible, it is preferable to have a macrostatistical description of a library, and a microstatistical description of a language. Accordingly, since this article is entitled "The Microstatistics of Text," we will be discussing primarily the problem of describing language with the aid of statistical analysis.

In view of the linguist's tendency to describe language in terms of word types and situation types rather than in terms of token counts, some effort must be made to justify microstatistical description. If the linguist's lack of quantitativeness were due solely to the laboriousness of making counts, the justification would be easy: we now have computers to do the counting.

However, the argument is more likely to be, "Why should one want to be quantitative about language, even if he has the means?" The linguist views language as a communicative facility and is interested in recording whatever is available for use in communication, regardless of how often it is used. Passing a test of plausibility or of recognizability by an informant is enough to establish an element of a language; given that this test is passed, the number of times the element actually comes into play in the speech of the populace or in its written material is of little relevance in the description of that language.

Since computers have entered the picture, however, there has not only been an increase in our ability to be quantitative about language, but also (ironically) an increase in the need to be quantitative. An example is machine translation, wherein the intrinsic post-editing burden is proportional to the number of

incorrect tokens. Therefore, in a functioning machine translation system we would not regard words like "sacroiliac" and "time" as being of equal interest; the latter may pass through the system at the rate of 1000 tokens a day, but the former might never occur. Sheer considerations of economy in such a case compel us to take the microstatistical viewpoint.

The justification of a microstatistical description is not only a question of programming economics, but also a question of establishing the best conceptual basis for discussions about computer handling of languages. There is evidence that we have been hampered in exploiting computers for processing language because of slow intellectual adaptation. The unfeeling use of predictive statistics has already been given as an example of this.

Another example which I have seen repeatedly in the literature of information retrieval is the argument: "We must relate terms in searching requests in order to guard against the possibility of retrieving documents about 'action by fluorinated plastics on bacteria' when we really want documents about 'action by bacteria on fluorinated plastics.'" When this situation is looked at with a quantitative eye, the problem disappears because usually only one of the two combinations will have any documents in correspondence with it, and this will usually be the combination of interest to searchers. Thus one can create for himself a non-existent problem by confusing what can happen in a retrieval system with what is likely to happen. Section II will give several more indications of how a microstatistical view can change our way of thinking about language handling problems.

II. Probability and Cognition in Text

The motive to think quantitatively about language goes beyond a desire to put counterexamplists to rout, or to talk in terms of probabilities rather than possibilities in designing language processing machines. The motive is that in the dynamics of people using language there is a relationship between the frequency of occurrence of a language element and the speed and manner of its interpretation by a listener or reader. It is well known that authors who use rare words are less likely to be understood than authors who use common words. The effect is pronounced: Katter (8) found that the ratio of comprehenders to non-comprehenders was six times as large for words having a Thorndike-Lorge frequency of 30 per million as for words having a frequency of one per million.

The effect should be even more dramatic if the probability spectrum were wider, and of course it is intrinsically much wider for word groups than for individual words. Consider, for example, adjacent-word pairs vis-à-vis single words. In regard to the latter, Zipf (2) shows us that a few words are highly probable in a short textual passage, having frequencies greater than one out

of a hundred tokens*, some are moderately probable, and the majority are somewhat improbable, occurring less than once in a million tokens of running text. But one could probably say that all dictionary words occur outside of the dictionary with some reasonable frequency, otherwise they wouldn't be in the dictionary. In other words the lower limit of the probability spectrum for single dictionary words could not be much below one billionth (10^{-9}).

With regard to the set of all possible adjacent-word pairs, however, one can feel safe in asserting that most of them never occur. Speaking in terms of probabilities, one could well have a spectrum all the way from $p = .01$ for a term like "of the" to $p = .00000000000000000004$, i.e., 4×10^{-21} , for a term like "vasoconstrictor Meistersinger." The latter number is computed from assumptions about the frequency of presentation of nonsense sequences in the world's literature and from the probability of picking any particular two words at random out of a 500,000 word dictionary. The probability spectrum could well extend lower than this, but the probability given above is already so low that if I had not said "vasoconstrictor Meistersinger" it probably never would have been said by anyone. (Of course, the probability of it occurring some day as part of a computer output may be somewhat higher.)

Suppose we compare, as an intellectual exercise, three groups of terms which are widely separated from each other on the probability spectrum, and ask what this probability separation means from a cognitive viewpoint. Consider the following three lists:

high school	high table	high escapes
black market	black chair	black Shanghai
nose cone	magnificent nose	powerless nose
foreign trade	foreign trains	foreign twenty

The three lists tie the words "high," "black," "nose," and "foreign" to other common words to form two-word structures. We would expect the structures on the left to be more common than most words in a 100,000 word dictionary. The middle list contains structures sufficiently infrequent that the reader might not be able to remember having encountered them in text. The structures at the right would be so rare that we might read for a lifetime without encountering one of them.

The lists differ from each other in another respect. When we hear a term on the left-hand list, we hear it as a unit and don't need to be conscious of its component words in order to understand it--it is practically equal to a word in the speed with which meaning is conveyed. The terms in the middle

*In most frequency counts, "the" will be most probable, making up 5% of the tokens, followed by "of" and "and," which occur about 3% of the time.

list are "descriptive," and we must be aware of the component-word meanings in order to "get the picture." In reading the term "high school" we do not have to think of the idea "height," but for "high table" we have to think of height because we do not recognize the term as a unit. The terms on the right-hand list are all somewhat meaningless and would have to be explained to the reader, either overtly or through context. Out of context they would convey little, and any sense we make out of them would be poetic rather than literal.

We assume that the understandability of a term is related to its probability of being used. Writers and speakers would be inclined toward two-word combinations such as those on the left, not only because these terms are more quickly understood, but also because they are more available in the speaker's own mental recall--so much so that it is a considerable mental strain to think up alternative ways of saying "high school" or "nose cone." Both the recognition and the recall mechanisms contribute to a reciprocal cause-and-effect relationship between the probability of being perceived and the probability of being uttered: if you hear something more often, you are more likely to say it--and if people say it more often, it is of course more often heard. Thus the brain reveals itself as a statistical machine which quite unwittingly counts, in its use of language; and the huge population of brains exchanging communications with each other in a modern society is an even more imposing statistical machine.

It is hard to study the relationship between the probability and the cognitive aspects of word groups, particularly at the lower end of the probability spectrum, where it is difficult to obtain information about actual probabilities. Several psycholinguists, however, have studied the upper end of the probability spectrum, and in particular Miller and Selfridge (9) have shown that nonsense sequences like

"....house to ask for is to earn our living
by working towards a goal for his team
in old New York was a wonderful place
wasn't it even pleasant to talk about
and laugh hard when he tells me lies
he should not tell me the reason why
you are is evident..."

are as easily recalled as passages of equivalent length lifted from meaningful text. What we have, in effect, in the above passage is an aggregation of highly probable five-word sequences. The probabilities could not be obtained directly, as by frequency counting text, but they could be obtained indirectly by allowing human subjects to fabricate five-word sequences in the natural process of sentence construction. The nonsense passage above was derived by giving one person the first four words, "house to ask for," allowing him to supply a fifth word by building a sentence like, "The house to ask for is that one on the corner," giving "to ask for is" to another person in order to get the next word, and in this manner continuing to give each successive person

the most recently generated four-word sequence so that he can supply a fifth word.

Though a complete frequency count of the entire English language literature might give a set of "most probable" fifth words at variance with the fifth words chosen by human subjects, there is no reason to believe that the result of the Miller and Selfridge experiment would thereby be changed. In other words, to reinforce a point made in Section I, it is not the exact numerical values of probabilities or other statistics we are interested in, but their orders of magnitude. This is particularly true in experiments relating language statistics and cognitive phenomena, because human minds are not likely to be sensitive to the second significant digit in the probability value of a word in a sequence--they may in many cases not even be sensitive to whether a word is the highest probable word in a sequence or the tenth highest probable in that sequence, as in a sequence like "he forgot to bring his ____."

statistics of adjacent words
as a function of dependency

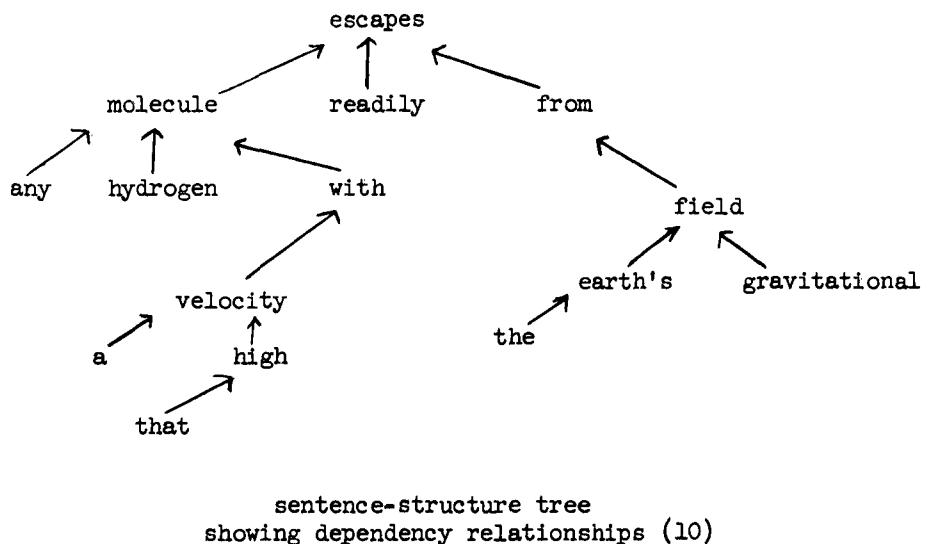
If an inventory of adjacent-word pairs in a large corpus were made, we would undoubtedly encounter in our inventory lists many strange bedfellows thrown together in the process of sentence-building. Improbable word pairs such as those listed on the right, above, are in fact generated as a consequence of tacking phrases together to form a sentence. For example, the improbable pair "high escapes" can be found in the sentence: "Any hydrogen molecule with a velocity that high escapes readily from the earth's gravitational field."

It is often said that grammatical analysis applied to automatic indexing can prevent such spurious word pairs from being accepted as indexing tags for a document. It is less widely realized that statistical analysis can do so also, and probably with less strain and less computer time: because the fact is that such spurious pairs tend not to repeat in text, whereas legitimate two-word terms do tend to repeat. Let us therefore explore this great and as yet untapped benefit of the probabilistic-cognitive relationship which we have been discussing.

In terms of one kind of grammatical analysis--dependency analysis--we would say that in the above sentence containing "high escapes" the word "high" is not dependent on "escapes" but on "velocity," i.e., because it actually modifies "velocity" and not "escapes." A dependency analysis would therefore separate "high" from "escapes" on the sentence-structure tree (see Figure 1). And so we would like to investigate the statistics of such dependency-separated pairs.

It is interesting, first, to inquire how often dependency-separated pairs occur in text. In my own inquiry on this I confined my attention to content words, as a result of my interest in automatic indexing; it is a simple matter in any

FIGURE 1



such automatic technique to include a list of the most common function words (the, of, at, his) in computer storage as "forbidden words" which we would categorically reject as potential indexing tags. I found in general that there was approximately one instance of a dependency-separated pair of content words for every fifty words of running text, and that for every instance of a dependency-separated pair there were about six instances of dependency-linked pairs, i.e., content-word pairs which would be directly linked on a sentence-structure tree.

Further, it seemed worthwhile to investigate the stability of the six-to-one ratio: did it vary from one sample of text to another and if so, was there any noticeable pattern to the variation? For each sample of text I counted content-word pairs until I had recorded 42 instances* of dependency-separated pairs, and I then computed the ratio: total content-word pairs/42. The data are as follows:

*The number 42 was chosen because that many lines of data were needed to fill one page in my notebook.

Text Sample

$$\text{Ratio} = \left(\frac{\text{total content-word pairs}}{\text{dependency-separated pairs}} \right)$$

<u>U. S. News & World Report</u> article on Indonesia	3.7
<u>Scientific American</u> article on magnetic resonance	5.3
<u>American Documentation</u> article on automatic indexing by author X (last half of article)	5.3
<u>Scientific American</u> article on beryllium	5.5
<u>ACM Journal</u> article on automatic indexing by author Y	5.7
Unpublished science fiction by author Y	5.7
<u>American Documentation</u> article on automatic indexing by author X (first half of same article)	6.0
<u>Psychological Abstracts</u>	6.1
<u>New York Times</u> article on deGaulle and the Common Market	6.1
<u>Scientific American</u> article on structural linguistics	6.7
Unpublished science fiction by author Y	6.9
<u>U. S. News & World Report</u> article on the steel industry	7.0
<u>American Documentation</u> article on automatic indexing by author Y (Sections I and II)	7.0
System Development Corporation abstracts of air defense system articles	7.1
Unpublished article on automatic indexing by author Y	7.5
<u>American Documentation</u> article on automatic indexing (Section III of same article by author Y)	7.9
<u>ACM Journal</u> article on structural linguistics	8.6

The ratios obtained had greater stability than I expected, and in the cases of the extreme values, 3.7 and 8.6, the causative factors seemed evident. For example, the article on Indonesia described Soviet military aid in great detail, with many phrases like "10 TU-16 jet bombers" and "four Skorii-class Russian destroyers." In dependency analysis, of course, the word "four" would be linked directly to "destroyers," which it modifies, and separated from "Skorii-class." Thus, the occurrence of very many such long terms creates an unusual number of dependency-separated pairs.

The 8.6 ratio for the article on structural linguistics seemed to arise from the presence of a great many two-word terms which were repeated from sentence to sentence in a tedious analysis, e.g., "The second pass links N_3 to V_2 in accordance with rule 5 from Table B."

If one ignores the articles with the 3.7 and 8.6 ratios as being "accountably unusual," one then notes that the spread of ratios for author Y is greater than the spread for articles not by author Y, which come from a variety of subject fields as well as authors. Thus it is not possible to say that the variation existing in the ratio is due to topic or to style.

The percent breakdown of the various types of dependency separation is also interesting. As already mentioned, most of the dependency separations come from nouns preceded by three or four modifiers. Out of the 714 cases tallied in the above table, 390 (54%) come from this source. (In three-word constructions like "finely decorated horse" or "stainless steel pot," no separations result in dependency analysis, because each word has a dependency link to one adjacent to it. Some cases always occur, of course, which are difficult to decide, like "European steel market." Is it "European steel" or "European market"? A transformational test plus context often enables one to decide, e.g., it is either "market for European steel" or "European market for steel." In the former case no dependency separation occurs, whereas in the latter "European" really modifies "market" and therefore the dependency analysis would separate it from "steel.")

The second largest category (18% of the 714 cases) was verbs occurring adjacent to noun modifiers, as in "the English alphabet contains 26 characters," or "all meticulous cosmologists persecute savage octogenarians." The next largest (9%) involved verbs preceded by prepositional phrase modifiers of the subject as in "the rate of growth rose to 5 percent." Had the sentence been "the growth rate rose to 5 percent," no dependency separation of that kind would have occurred.

The fourth largest (5%) was nouns (subject or object) followed by adverbs, as in "the analysis revealed a pattern strongly resembling a language" or "the plane suddenly hit a downdraft." Next (4%) was prepositional phrases followed by nouns or modified nouns (this is a type commonly brought up in language discussions by counterexamplists), such as "in solving such a problem children from the experimental group excelled those from the control group." The remaining miscellaneous 10% separated such juxtapositions as direct and indirect objects, terminal words in restrictive clauses and what follows, and words brought together as a result of omissions of understood parts of speech ("this is not the way people should behave").

The most direct way to test the assertion that such dependency-separable pairs tend not to repeat is to inspect, on a token-by-token basis, all instances of word pairs which occur more than a given number of times in a large corpus. I chose to investigate a corpus of 45,000 words comprising some 618 abstracts from Psychological Abstracts. I looked up in the text all

tokens of pairs which occurred five or more times.* I hoped to find not only that there were few or no repetitions, but that if there were repetitions some statistical way would be available to prevent error from occurring in some automatic term-selecting system. I also hoped to find very few instances of a dependency-separated pair token just happening to be identical with a pair which is not normally separated in dependency analysis.

An example of the latter is: "In seeking a higher status in the West Germany has exercised great restraint." Editors will tell you that one should not construct such sentences as this, but this doesn't stop the counterexamplists.

The number of pairs occurring five or more times in the corpus proved to be 58, amounting to 413 pair tokens. In the process of inspecting these tokens for dependence, some additional information was gathered about the immediate environment of the pairs, and this is given in the table on the following page. Liberties were taken in equating singular, plural, and participial forms, based on the knowledge of the existence of workable suffix-splitting programs (11); no particular analytical gain obtains from working on a pure identical form basis.

The first thing of interest about the data of Table I, in the light of our discussion at the beginning of Section II, is that most of the pairs would clearly be of the type "black market," "nose cone," etc., word groups which tend to be perceived as a unit. We see "social status," "intelligence test," and "New York" as terms we recognize and have seen or heard many times. Furthermore, such terms as "factor analysis," "control group," "case history," and "group therapy," are pillars of the psychologist's jargon. Even the one case of a dependency-separable pair, "central nervous," is part of a three-word label ("central nervous system") so familiar to life scientists that many people use its abbreviation CNS.

On the other hand, descriptive combinations are in the minority, and many of these to a psychologist are so familiar that recognition as a unit must often occur. "Results indicate" may be descriptive to some readers, requiring words to be comprehended individually, but a psychologist--especially one in the habit of perusing abstracts--may absorb the two words as "one blob" of meaning.

Most pleasing of all, a psychologist would find no pairs in Table I which, like

*This was made possible through the use of an IBM 7090 text inventory program (6) which counts, lists alphabetically, and indexes all content words and word pairs in a corpus of any size. The original program was written by John Olney of the System Development Corporation, and the indexing feature was added by Keren McConlogue also of System Development Corporation.

TABLE I

<u>Word Pair</u>	<u>Frequency</u>	<u>Finding</u> (code key below)	<u>Word Pair</u>	<u>Frequency</u>	<u>Finding</u>
Mental ability(ies)	6	OK	Results indicate(d)	7	OK
Academic achievement	5	OK	Glucose injection(s)	5	OK
Reading achievement	5	OK	Brain injury	6	OK + S _a
Factor analysis(es)	5	OK + N	Head injury(ies)	5	OK + N
Author analyzes	6	OK	Central nervous	8	8T
Occupational aspiration(s)	5	OK	Cerebral palsy(ied)	15	OK + N
Test battery	7	OK + 4M (2,2)	Counseling psychology	5	OK
Human behavior	5	OK	Group psychotherapy	8	OK
Social change(s)	6	OK	Case report(s)	6	OK
Deaf child(ren)	6	OK	Mentally retarded	8	OK + 5N (5)
Handicapped child(ren)	9	OK + 6M (4)	Chronic schizophrenic(s)	5	OK + 4N (2)
Retarded child(ren)	7	OK + 5M (5)	High school(s)	19	OK + 13N (6)
Occupational choice(s)	5	OK + N	Public school(s)	6	OK + 4N (2)
Brain damage	5	OK	Social science(s)	9	OK + 4N
Present day	6	OK + 5N	Scale score	5	OK + 5M (3)
Author describes	7	OK	Test scores	5	OK + 4M (2)
Significant difference(s)	14	OK + 3M (3) + S _b	Statistically signifi- cant	5	OK + 3N (3)
Short form(s)	5	OK	Normal Ss	5	OK
Call girl(s)	6	OK	Social status	6	OK
Temporal lobe	5	OK + N	College student(s)	14	OK
Psychosomatic medicine	5	OK	School students	7	OK + 7M (6)
Experimental group(s)	9	OK	Case study(ies)	6	OK
Control group(s)	9	OK	Nervous system(s)	10	OK + 2N
Mentally handicapped	6	OK + 5N (4) + S _a	Intelligence test	7	OK + 4N (2)
Mental health	9	OK + 4N	Projective test(s)	7	OK + 2N (S _a)
Case history(ies)	15	OK	Group therapy(ist)	7	OK + N
Mental hygiene	7	OK + 4N (2)	Social worker(s)	7	OK
Ambient illumination	5	OK	First year	6	OK + N + S _a
Wage incentive(s)	7	OK + 2N	New York	7	OK + 3N (2) + S _a

Codes OK = None of the pair tokens are separated in dependency analysis.

T = Tokens of this pair are separated (number preceding T shows number of pair tokens broken).

xN = This pair modifies a noun in x of the cases (numbers in parentheses show how many of the x nouns are identical to each other).

xM = This pair has the first word modified in x of the cases (numbers in parentheses show how many of the modifiers are identical to each other).

S_a = At least one pair token (usually only one) followed by word which is dependency-separated from either word of the pair.

S_b = Same as S_a for word preceding the pair.

"high escapes," would not be readily interpretable.* It is perhaps not surprising that our inventory should reveal these things; and yet truths of the type shown in Table I never seem to wend their way into linguistic discussions. This is why it was suggested in Section I that statistical awareness can change our conceptual approach to language. And now that fast and capacious computers are available, even the unimaginative can acquire this statistical awareness.

The solid presence of the code OK for all the pairs but one in Table I is an impressive confirmation of our original hopes. The lone violation of our expectations, "central nervous," is redeemable in any automatic term-selecting process because of the repetition of the word "system" immediately following. In fact, in this example it follows in all eight occurrences of "central nervous." But have we been lucky? What if "system" had followed only three or four times? There are two answers to this question. One is that if "central nervous" had occurred several times not followed by some content word, it probably would be a semantic entity in its own right.

The second answer is that even if the word following occurs only part of the time, we might accept it as part of the term anyway, as a safety measure. At best it could salvage the term, and at least it probably would form an understandable descriptive addition--its repetition giving us a reasonable guarantee that it isn't dependency-separable from the last word of the pair. It was to shed some light on this very situation that I gathered the additional data in Table I, shown in the codes immediately to the right of the "OK column."

The additional data were collected to assess the likelihood of two kinds of error: 1) leaving out a word adjacent to a pair when this word would contribute substantially to the pair's meaning (e.g., "school students" requires "high" or "public" preceding it to assure clarity of meaning) and 2) including a word adjacent to a pair when the word contributes only semantic static (e.g., "brain injury covers" in the sentence "This survey of research on the effects of brain injury covers studies done...,etc.").

It is to be noted that to protect oneself against an error of one kind is to expose oneself to an error of the other kind. It is also to be noted that errors of the first kind usually involve leaving out words which have a dependency link with the closest word of the pair (I'm sure a top-flight counter-examplist could think up an exception), and errors of the second involve including

*When the list was shown to a psychologist, he recognized all the terms but "call girl," "first year," and "New York" as constructions he had seen in the psychological literature. When questioned if he had seen the three exceptions in non-psychological text, he said, "Of course."

a word not having a dependency link with either word of the pair.* Keeping track of words of both these kinds leaves out the most common kind of word adjacent to the pair, and that is the first modifier of a doubly modified noun (e.g., "superior reading achievement," "general mental ability"). I left these words out of the survey primarily because they seldom seemed to contribute to either kind of error, and it makes little difference in the semantic effectiveness of a pair as an index tag whether or not we add such words to the pair.

Turning now to the codes of Table I, cases of type M ("statistically significant differences") and type N ("mental hygiene workshop") are of the meaning-contributing variety, and we hope to include the added word. Cases of type S_a ("brain injury covers") and type S_b ("high escapes readily") are of the static-injecting variety, and we hope these words are not included.

Table I reveals 105 occurrences of words of types M and N, of which 55 have repetitions; in contrast, only eight occurrences of types S_a and S_b were found, and there were no repetitions. A list of the pairs with adjacent words which repeat is given below, in the same order as their corresponding codes are listed in Table I. Note that some "triplets" occur twice in Table I, once as a pair followed by an N-type word and again as a pair preceded by an M-type word. The N or M words are underlined. The number of occurrences of the three-word structures is also given.

Table I (Supplement)

<u>Personality test battery</u>	2	<u>High school student(s)</u>	6
<u>Aptitude test battery</u>	2	<u>Public school boys</u>	2
<u>Mentally handicapped children</u>	4	<u>Full scale score</u>	3
<u>Mentally retarded children</u>	5	<u>Intelligence test scores</u>	2
<u>Statistically significant differences</u>	3	<u>Statistically significant</u>	
<u>Mentally handicapped children</u>	4	<u>differences</u>	3
<u>Mental hygiene workshop</u>	2	<u>High school student(s)</u>	6
<u>Mentally retarded children</u>	5	<u>Intelligence test scores</u>	2
<u>Chronic schizophrenic patients</u>	2	<u>New York City</u>	2

Thus it would appear, from this particular corpus at least, that if we follow a policy of accepting all words adjacent to a pair when they occur there two or more times, we make errors of the first kind only for three-word structures which occur once, and we practically never make errors of the second kind. And of course our total take of semantically effective three-word structures will be much larger than what is listed above, consisting mostly of doubly modified nouns which, for reasons explained above, were not coded in Table I. For corpora much larger than our 45,000 words the policy might not be a safe one--sooner or later the laws of chance if nothing else will result in

*This definition was chosen to exclude one kind of dependency-separable configuration, i.e., it was found that whenever a word preceding a pair had a dependency link with the final word of the pair (as in "designs retrieval systems"), the inclusion of the word practically never produced a term difficult to interpret.

TABLE II

<u>Dependency-Separated Pair</u>	<u>Type of Separation</u>	<u>Frequency</u>	<u>Next Dependency-Linked Pair</u>	<u>Frequency</u>
Freud's revolutionary	T	1	Contribution consists	1
Motivation presupposes	PV	1	Teleological attitude	1
Freud's grandiose	T	1	Valid instinct-model	1
Format completely	NB	1	Five parts	1
20 different	T	1	Psychoanalytic therapy	1
American Psychoanalytic	T	1	General theory	3
Writings especially	NB	1	Generalized systems	1
Analysis computational	T ₂	1	Forced choice	2
Simple random	T	1	Uncertain outcomes	1
Binomial probability	T	1	Statistical inference	1
Drawing statistical	VT	1	Arithmetic mean	1
Two quantitative	T	1	Efficient estimation	1
Non-differentiating	T	1	Items suggesting	1
General mental	T	1	Biographical dictionary	1
Development due	PM	1	Well-integrated personality	1
Rabbit showed	PV	1	Significant rises	1
Rate following	PM	1	Areas responding	1
Following glucose	VT	1	Responding differentially	1
Different bodily	T	1	Author suggests	2
Intravenous glucose	T	3	Reinforcement process	1
Primary reinforcing	T	1	Nutritive need	1
Important cognitive	T	1	Briefly summarize	1
Nuclei following	PM	1	38 cats	1
Following selective	VT	1	Thalamic nuclei	2
Selective cortical	T	1	Three groups	3
Degehnration following	OM	1	Second group	3
Following circumscribed	VT	1	Third group	1
Circumscribed local	T	1	Secondary degeneration	1
Following extensive	VT	1	3 groups	1
Normal human	T	1	Comparative material	1
Book deals	PV	2	Civilized mind	1
Attitude perception	PN	1	Action attitude	4
Subjective salt	T	1	Indirect evidence	1
Multidimensional subjective	T	1	Genetic approach	1
Gravity shows	PV	1	Inverse tendency	1
Obtain better	VT	1	Different ways	1
Internal stimulus	T	1	Women react	1
React more	VB	1	Perceptual development	1
Men actively	V	1	Sensory-tonic field	1
Explains perceptual	VT ^o	1	Prism vergence	2
Negative visual	T	1	20 neurotics	2
Using 106	VT	1	20 psychotics	2

Code Key for Table II

<u>Codes</u>	T = First pair of a 3-word term or doubly modified noun.	PV = Prepositional phrase modifying subject followed by verb.
T ₂	Second pair in a 4-word term.	PM = Prepositional phrase followed by phrase modifying the verb.
VT	Verb followed by term or modified noun.	(e.g., "X showed rises in
VB	Verb followed by double adverb.	temperature following
V _o	Verb omitted in parallel construction.	exercise.")
NB	Noun followed by adverb.	PN = Initial prepositional phrase
OM	Object followed by phrase modifying the verb.	followed by subject.

repetitions of S_a or S_b cases. Whether or how much the frequency threshold will have to be raised, only further investigation can answer.

Some readers may be unsatisfied with the data of Table I because it deals only with word pairs occurring five or more times; they may think, "But what about word pairs which occur twice?" Of course, dealing with all word pairs occurring more than once would have been an exhausting task. However, there is a way of getting a worm's eye view of this part of the universe: in the analysis leading to Table I the question was put, "Do word pairs which repeat ever involve dependency-separated tokens?"; but we can also conduct an analysis based on the question, "Do dependency-separated tokens ever repeat?"

To answer this question I plowed through Psychological Abstracts looking for dependency-separated pairs. Whenever one was found I looked it up in the alphabetical inventory listing (6) for the entire corpus, and noted its frequency. Then, starting from the point in text where the dependency-separated pair occurred, I looked for the next dependency-linked pair which did not have a word in common with the dependency-separated pair. Looking up the frequency of this pair also, I could make a direct comparison of the statistics of dependency-separated pairs and dependency-linked pairs. This comparison is shown in Table II.

Two repeating dependency-separated pairs were found, "intravenous glucose" and "book deals." "Intravenous glucose" was actually the first two words of "intravenous glucose injection" in all three cases. This case is of course the same as the imperfection in Table I, "central nervous," and is dealt with by the adjacent-word inclusion policy we have already discussed. "Book deals," however, is the nasty one I hoped wouldn't happen--where a pair that is independent, as in the sentence, "The remainder of this book deals with X," just happens to be identical to a pair with a dependency link, as in the sentence, "This book deals with Y." Of course, this kind of thing has to happen sometime; otherwise the aura of plausibility so advantageous to the counterexamplist would be absent.

statistics of phrases

Up to this point we have confined our attention to the statistics of adjacent content words in groups of two and three. It would be interesting to know if the principles illustrated in Tables I and II might be even more general than has been shown so far. A broader question for us might be, "Do repeating word groups of any length ever omit nodes in a dependency tree?" What we are suggesting is that phrases which hang together as cognitive entities in people's minds, and which therefore are likely to repeat, also hang together in a grammatical way, i.e., exist as integral substructures in a sentence-structure tree like that in Figure 1.

Considering the sentence depicted in Figure 1, one notes that it contains 11 possible strings of five successive words, as would be true of any 15-word sentence. Checking the position of each of these on the dependency tree, one further notes that only three of the strings form integral substructures, namely, "hydrogen molecule with a velocity," "with a velocity that high," and "from the earth's gravitational field." "Any hydrogen molecule with a," for example, would not form an integral substructure because, though the first four words are connected together by dependency links, the article "a" is isolated. "Velocity" has to be introduced into the network to make it "all one piece," i.e., integral. "Velocity," then is an omitted node in the sub-structure formed by the first five words in the sentence.

In investigating the question of whether repeating word groups ever omit nodes, I was again influenced in my procedure by a problem of long-standing interest to me, that of picking content-indicative word groups out of text. The emphasis was therefore to be on content words, function words being valued chiefly for their functional capability--to relate content words. My procedure was also influenced by a desire to encourage word groups to repeat themselves, since word group repetition was what I wanted to study. So I confined myself to word structures all of which had one content word in common, the word "computer," and all of which were written by author Y in a corpus of some 110,000 words. The fact of sameness of authorship has the additional advantage of loading the dice in favor of having node-omitting word groups repeat, as a result of the sentence-building habits that a person has, so that if the results of the investigation are good with a corpus authored by one person, they should be even better for a corpus of similar size authored by many people.

The procedure was to start at the beginning of the corpus, to note all occurrences of the word "computer," and to record in each case the string of words to the left and to the right so as to include either

- 1) the entirety of any string of content words unbroken by a function word, or

- 2) any string of function words adjacent to "computer" plus the adjoined content word or string of content words, or
- 3) any single content word adjacent to "computer," plus the adjoining string of function words, plus the adjoining content word or string of content words.

In all cases the recorded strings began and ended with content words, and in no case was a string followed across a punctuation mark such as a period or comma; this did not apply, of course, to such punctuation marks as hyphens or apostrophes. Since the corpus had not been keypunched, and since no computer program would have been available to carry out the above procedure anyway, the entire chore had to be performed by hand. Strings were filed under their component content words, and repetitions were kept track of by tallying. In the process a piece of paper 30" by 22" was pretty nearly filled up, since there were a large number of occurrences of the word "computer."

The results of the tally are shown in Table III. The first thing we note is that the repeating word groups are generally short; on a token basis the majority of them are just content-word pairs. This is not surprising, although one should be careful not to say that this is expectable "by chance." Selection of words in communication can hardly be mistaken for a random process, though more than once people have asked me to prove by statistical means that it was indeed not a random process. Anyone, however, who has observed case after case such as the one involving the word "computer" and the word "program" occurring together in sentences 52 times and just happening to occur adjacently 35 of those times, with the word "computer" always preceding "program," knows how ludicrous such a request is.

The second thing which Table III reveals is tendency of the omitted nodes to occur more often among the less frequent phrases. Among the 75 tokens of phrases of low frequency (two or three occurrences) we find 19% of the tokens involving omitted nodes. Among the 80 tokens of phrases of intermediate frequency (four to ten occurrences) the percentage diminishes to 11%. In the 75 tokens of the four phrases of highest frequency there are no omitted nodes at all. This trend reinforces our original hunch, above, that "...phrases which hang together as cognitive entities in people's minds, and which therefore are more likely to repeat, also hang together in a grammatical way, i.e., are integral substructures in a sentence-structure tree..."

Next we compare the results in Table I to those in Table III; the omitted nodes in the latter are more numerous and more widely distributed than those in the former. This is partly a consequence of the generation of all of the phrases in Table III by one author, and partly a consequence of including all tokens with a frequency greater than one (Table I includes only tokens with frequencies of 5 or more). We recall that in Table I only one out of 58 word-pair types involved omitted nodes, i.e., "central nervous," but since all 8 tokens involved the same node, "system," we could restore the integrated

TABLE III

Phrase	Frequency	Phrase Types	Errors	Phrase	Frequency	Phrase Types	Errors
C program(s,ing)	35	T		entries beginning with "C"	3	NVPT**	
C program was written	2	TVA		entries beginning with "C program(ming)"	2	NVPT	
C analysis	17	T		C language handling	3	T**	1
C analysis of text	7	TPN		C language-handling systems	2	T	
C analysis of library(ies)	3	TPN, TPT*	1	use(ing) a C	3	VN	
C storage	12	T		C age	2	T	
placed in C storage	2	VPT		aid of C(s)	2	NPN, NPT*	1
C-based	11	T		1 C application(s)	2	T	
C-based retrieval	4	T****		C capacity	2	T	
C-based retrieval systems	3	T		2 Western Joint C Conference	2	T	
C-based library	2	T**		1 impact of C(s)	2	NPN, NPT*	1
C handle(ing)	10	T, V _o *NV		1 C-implementable	2	T	
C handling of language	9	TPN, TPT*		Cs in placing an			
digital C(s)	9	T		individual	2	V**NPVN	2
general-purpose digital C(s)	3	T		C instructions	2	T	
C time	8	T		language on a C	2	V**NPN	2
use(ing) C(s)	7	VN, VT**		C memory	2	T	
use C programs	2	VP		role of the C	2	NPN	
use(ing) the C	6	VN		C systems	2	T	
general-purpose C(s)	6	T		tomorrow's Cs	2	T	
Cized library	5	T*		C usage	2	T	
use of Cs	5	NPN		C use	2	T	
C run(s)	4	T		use of a C	2	NPN	
application of Cs	3	MPN		C(s) should be used	2	NV _o VA	
C assistance	3	T					
electronic Cs	3	T					

Part-of-Speech Code

N = Noun
 T = Term or modified noun
 V = Verb
 V_o = Auxiliary verb or first element of a complex verb
 A = Adjective or participle
 P = Preposition
 (Articles not coded)

(C = computer, Cs = computers)

Omission Code

An asterisk to the right of a part-of-speech code symbol indicates the location of an omitted node. Thus TPT* means the omitted node is the terminal noun of T. In the case of the first entry of an asterisk in the above table, this noun was "structure" in "computer analysis of library structure." Similarly, V_o*NV refers to the passage "let the computer handle the situation," where "let" is the omitted node. Multiple asterisks show the number of phrase tokens involving that kind of node omission. Number in "Error" column refers to number of node omissions which cannot be restored by statistical means.

Token subsets are listed immediately following the parent set. (Example: The 7 "C analysis of text" tokens are included in the 17 "C analysis" tokens.)

substructure by having a program which keeps track of what lies on either side of a given word pair as its tokens are inspected, and which adopts* as a third word in the term any word recurring in positions adjacent to the pair. Table III is set up in order to illustrate the restoration of omitted nodes by such a program. As a typical example, note the entry "C-based retrieval," whose Phrase Type entry includes four asterisks. Immediately following is the entry "C-based retrieval systems," which accounts for three of the four occurrences of "C-based retrieval," and which shows that the word "systems" is restorable as the missing node as a result of its triple occurrence. Three other similar pairs of entries can be found in Table III by comparing the asterisk counts to the number in the Error column. Altogether, nine out of the 23 omitted nodes are restored.

We've already explored the workability of this restoration procedure, in effect, when (following the discussion of Table I) we tested it with the 57 pair types which did not involve omitted nodes, and the success of the test in that case meant merely that no new omitted nodes became generated as a result of adding a third word to the pair. The test "worked," in that none of the S_a or S_b type additions recurred, and ergo when an added word did recur, no omitted nodes were thereby introduced.

In Table III each asterisk is actually a general case of S_a or S_b , and we are of course disturbed to see several cases of double, triple, and quadruple asterisks. However, when we pay close attention to the pattern of restoration, we find a situation which is fairly compatible with what was observed about the Table I data: the single asterisks are not restored, the double asterisks are sometimes restored, and the triple and quadruple asterisks are restored all but one; this leaves the Error column clean of 3's and 4's, but there are still some 2's.

So let's look more closely at what happened to the double asterisks. Note that two kinds of cases are present, double asterisks where the frequency of the phrase was 2 and double asterisks where the frequency of the phrase was greater than 2. Two out of three of the 2's in the Error column came from double asterisks of the former type, whereas when the frequency of the phrase was greater than 2, the double asterisks were "restored" in two out of three cases.

This leaves "C-based library" as the only phrase which is not in accord with the Table I analysis. It is to be remembered that the pairs in Table I all had a frequency of five or greater, and therefore our restoration procedure

*From an automatic term-selection viewpoint, adopting a third word does not necessarily mean throwing out the original word pair per se; whether or not we abolish the pair would depend partly on what fraction of the pair tokens were supplemented by the third word.

never had a chance to operate on pairs which occurred twice, as in the Table III data. If we had included pairs occurring twice, we might well have seen some recurring S_a 's or S_b 's.

And so, by applying the restoration procedure under the more stringent conditions of Table III, we have not necessarily made the procedure look worse, we have merely sharpened the frequency boundary between phrases we can safely accept and those we might better reject--for corpora of these sizes. Now, instead of rejecting all things which occur once, we reject in addition all things which occur twice without having been derived from phrases of higher frequency via the node-restoring procedure (remembering that the procedure doesn't necessarily actually have to restore in the course of the derivation).

Such a rule loses for us most of the terms on the right-hand side of Table III, namely, those from "C age" on down, and this may seem like a considerable loss, just to avoid accepting an occasional unintegrated substructure. Our policy in such a decision depends greatly on our purposes. If we are indexing a document, there will usually be enough phrases occurring three or more times so that the loss of twice-occurring phrases is no loss at all. On the other hand, if we are automatically producing a thesaurus, we may want every term or phrase we can get, even at the risk of producing an occasional lemon.

The remarkable thing about this whole business is that here we are giving enormous importance to the number 2--should we honor it or not? To the usual statistical mentality, imbued with vague fears of small samples, such thinking might seem as alien as alchemy. Yet, from an inductive viewpoint it is not at all unreasonable: if something improbable happens once, well--improbable things do happen; if it happens twice, it could still be a fluke; but if it happens three times, we can no longer consider it improbable.

III. Statistical Operations in Indexing and Abstracting

I was led to the studies of Section II mainly because of my interest in automatic indexing and the related problem of condensed representation. I have often felt that the "linguistic barrier" could be circumvented in some text-processing applications through use of the most elementary kind of statistics. However, though in previous articles (7, 12) I have described text-processing systems whose output often took the form of terms or phrases, I was not able to state in detail how these groups of words might be derived.

In one of the concluding sentences of the more recent article (7), I said: "...we have not yet talked about frequencies of terms or phrase-length constructions, which a combination of statistical and structural linguistic techniques should eventually enable us to deal with...." It now appears that there is a great deal along this line which could be done by statistical technique alone. And it must be kept in mind that Section II doesn't even begin to expose the possibilities of unaided statistical technique--for

example, one could rescue many a doubly or singly occurring term or phrase in a document by checking a listing derived from the parent corpus; or in a corpus by checking against another corpus. One can envisage bank after bank of reference glossaries with entries untouched by the labored hand of the linguist.

One of the specific outputs I've been concerned with is an association map (7,12), which diagrammatically links pairs of words which co-occur in the same documents (or other unit of text) with unusually high frequency. As often as not, the high co-occurrence is a result of the repetition of some term or phrase; if such is the case, we would certainly like to show it on the map. Even if the co-occurrence is not due primarily to term or phrase repetition, we might still want to show two associatively linked words in a grammatically connected way, if only to improve the interpretability of the map. There are pros and cons to the latter viewpoint, but unless we have the capability to pick the terms and phrases out of text, these pros and cons are academic.

Taking the 110,000 word corpus used in Section II, I selected the 14 most frequent words with the intention of finding which of the 91 possible word pairs were most strongly associated for presence in the same sentence, determining which of the pairs are parts of repeating word groups, and correlating the two. Here again, absence of appropriate computer programs made it necessary to go through the entire corpus manually (there is much to be said for this sort of thing if it is to be a one-shot process). The 30 most strongly associated pairs are listed in Table IV; their most frequent grammatical relationships are shown; the frequency thereof is given, and the involvement of omitted nodes is indicated.

Our hopes are answered affirmatively by Table IV. For most of the strongly associated words there exists some reliable syntactic relationship. For those word groups with a frequency of five or greater there exists semantic as well as syntactic reliability; that is, the apparent meaning of the words as elements of the word groups is the same as it is generally throughout the corpus. At frequencies lower than five this is often not true, one interesting example being "programming language," in which the meaning of "language" is different from that in its usual use in the corpus, i.e., a programming language is not a natural language. In principle, this same thing could happen in the cases of word groups with a frequency greater than five, although our microstatistical mode of thinking cautions us to await data before bewailing the possibility.

Another encouraging thing about this data is that every one of the 14 most frequent words in the corpus has two or more syntactic combinations with a frequency of five or greater (some are not shown in Table IV because the parent word pair is not in the top 30 in association strength). This is good news for any kind of an automatic indexing procedure, and especially for the associative type of index, where sterile word-connecting links can now be replaced by indicators of syntactic relationship.

TABLE IV

Strength of Association	Frequent Groups	Frequency	Strength of Association	Frequent Groups	Frequency
.328	Information retrieval	120	.107	Text compiler program	2
.196	Machine searching	34	.104	Computer analysis of text	7
.193	Words in <u> </u> text	24	.097	Indexing word	21
	Words in text	8	.096	Word "program"	3
.189	Retrieval system(s)	55	.092	Documented information	15
.187	Number of <u> </u> words	19	.090	(no proximity)	
	Number of word <u>*****</u>	5	.089	(no proximity)	
.154	Document retrieval	10	.088	Information search(ing)	14
	Retrieval of <u> </u> document(s)	8	.084	Information processing	
.146	Computer program(s, ming)	35		system	1
.126	Computer handling of language	9	.080	Searching of text	7
.125	Index searching	2		Search <u> </u> text	7
.124	Word(s) in <u> </u> document	4		Text searching	7
.121	Search document <u> </u> *	1	.079	Number of text <u> </u> <u> </u> *	1
.120	Programming language(s)	3	.073	Words in <u> </u> language	4
.115	Searching system(s)	23	.073	Computer in information <u> </u> *	1
.115	Computerized library*	5	.070	Language one text (*)	1
.112	Document(s) in <u> </u> library	6	.067	Text retrieval	5

Explanation of Table:

1. Strength of association is computed from the function $f/(a+b-f)$, where a and b are the frequencies of occurrence of words a and b in the corpus and f is the frequency of co-occurrence of these words in the same sentence.
2. Blanks in the word groups stand for single words, which could be either function words or content words. The data were derived according to rules which could be executed by a computer, and one rule was: when a number of structures involving words a and b are all the same with the exception of one word, they can be tallied together. This rule substantially increases the tally in some cases.
3. Asterisks refer to omitted nodes, as in Table III. Note that only in one case, above, can a restoration procedure recover the node, i.e., "number of word ," which in three cases is "number of word pairs." The symbol (*) refers to an omitted node which is remotely positioned in the sentence.
4. The rules used in deriving the data attempted to find all repeating word structures with a frequency of five or greater. When no such word structure is found, structures with smaller frequencies are accepted, provided they obey a rule of proximity, i.e., that words a and b should be separated by no more than two intervening words. Where no such structure is found, the designation "no proximity" was entered for that word pair.

But the most interesting thing of all, to me at least, about the Table IV data is the existence of what might be called the "pinch effect" in sentence structure. The top 30 word pairs in association strength include all of the groups with a frequency of eight or greater, and all but one of the groups with a frequency of seven. Since the median value of f in the association function $f/(a+b-f)$ is 18 for the 61 pairs of smallest association, there is no reason why word groups of frequency in the range 8-15 shouldn't be scattered through almost the entire spectrum of association strengths; but they are not--they are decisively concentrated in the top 30 values.

This leads us to the tentative conclusion that words which are strongly associated cognitively, and hence which find their way into the same sentence with unusual frequency, are also apt to be thrown into a close grammatical relationship. It is difficult to criticize this reasoning as putting the cart before the horse, in a causational sense, because only five of the word groups shown in Table IV are identifiable as major factors in the high co-occurrence. Moreover, it does not add up to an unreasonable picture of human use of language: the need to make frequent simultaneous reference to two concepts can well cause one to try to incorporate those concepts into some term or phrase-like structure, for the sake of faster retrieval in speaking.

statistics and automatic abstracting

H. P. Luhn's choice of word frequency as a basic principle in automatic abstracting undoubtedly had good intuition behind it. In fact, it has sometimes seemed to me that he hit upon the one approach which is unavoidable in the abstracting problem. Regardless of whatever other techniques may eventually be used in conjunction, frequency counting seems certain to be a mainstay. The essence of abstracting is description, and as we pointed out in Section I, the essence of describing something which is much larger than the recognizable elements of which it is composed is counting and summing up of those elements.

If abstracting is to be a purely automatic procedure, it must rely on rules which the undependable people who write the documents cannot thwart. One approach in current vogue, the "pragmatic approach," in which machines search for author-implanted cues such as "this article will discuss" or "summarizing our results," is an excellent example of a complex of rules which an individual author can thwart: there is no law which says that the author has to use the cues in the computer's glossary, or for that matter any cues at all. This is not to say that for this reason we shouldn't investigate the pragmatic approach, which after all may work well in conjunction with frequency counting. This is only to point out that procedure may fail if it depends on behaviors which it is easy for authors to avoid.

By the same token, a procedure which depends on behaviors which are hard to avoid can be expected to work reliably. One behavior which is hard to avoid is repeating words which pertain to what one is talking about; the longer the article and the narrower the topic, the more difficult it is to avoid such repetition. And, of course, frequency counts will select these words which it is difficult to avoid repeating. The perversity of human nature which makes it possible to avoid pragmatic cues can have little effect in avoiding repetition--even among journalists who pride themselves on their ability not to use the same word twice. It has been rumored that Time magazine, at the apex of journalistic style, has this non-repetitive characteristic; I urge anyone who believes this to inspect a moderately long Time article. Avoidance of repetition in a long article would take an intellectual effort an order of magnitude greater than the usual labor in writing an article.

Granted that repetition is difficult to avoid, it is still possible for an author to repeat the "wrong words," i.e., words which will give a misleading impression of what an article is about. My feeling, based on the making of numerous spot counts in various fields, is that this is possible in some fields but extremely difficult in others. In fields such as artificial intelligence and system science, where one commonly defines his own terms, the mechanism for "wrong word repetition" is evident, but in fields such as chemistry and biology, with established vocabularies, an author is hard put to avoid repeating the words which, out of context, would indicate more or less exactly what he is talking about. Added insurance of the accuracy of this indication might come from automatic procedure for syntactically relating the most frequent words, which the data of Table IV would seem to show feasible.

Let's take a close look at another non-statistical approach to automatic abstracting, the use of syntax analysis in condensing articles as described by Climenson, Hardwick, and Jacobson (13). These authors experimented with the idea of article condensation through the elimination of modifiers, prepositional phrases, dependent clauses, and other subordinate sentence elements, thereby retaining only the basic skeleton of each sentence, usually the unmodified subject and predicate, but occasionally retaining also the skeleton of a long and important independent clause or an important modifying element. Thus, the sentence in Figure 1 might be condensed from "Any hydrogen molecule with a velocity that high escapes readily from the earth's gravitational field" to "Molecule escapes from gravitational field." In essence, the lower branches of the dependency tree have been pruned off.

Climenson, et al, conclude with a description of three ways of combining syntax analysis with the Luhn-type statistical method. Of course, going into any further detail on the combination of syntactic and statistical methods was not within the scope of their paper, but it is within the scope of our discussion here to inquire whether or not hidden perils exist in the syntactic approach which are engendered by variation in author behaviors. For example, is it true, (as Climenson, et al, maintain) that "the main information content of the sentence" is retained after the syntactic pruning.

One could challenge their statement in true counterexamplist fashion by constructing a sentence like, "I think that it may very well be true that you will be elected," where the meat of the sentence, "you will be elected," would be discarded. Microstatistical thinking, however, would ask for an investigation of the actual prevalence of such sentences.

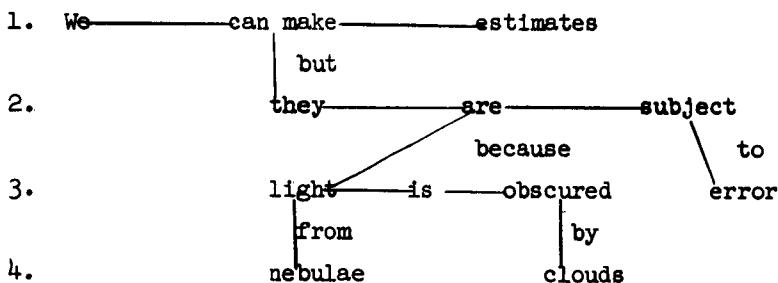
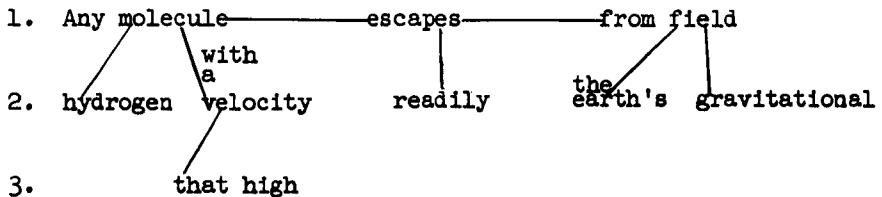
Before such an investigation is started, one has to decide on objective procedures for determining which words constitute the "meat of the sentence." Since we have already assumed that there is something important about words which an author chooses to use repeatedly, it seems natural as a first approximation to investigate the extent to which the most frequent words in a document are dropped. Some people might object to this on information-theoretical grounds, contending that the most frequent words contain the least information. However, this objection can be met squarely by drawing a distinction between "information carried by the word in context" and "information about the document." Earlier in this paper the term "vasoconstrictor Meistersinger" was used. Any information theorist would agree that this term contains an enormous amount of information, but it is not information about this paper, and certainly not something we would like to see in the abstract of this paper.

Accordingly, I chose two articles, an article on automatic indexing by author Y (see Section II) and a Scientific American article on radio astronomy, and chose for investigation those sentences which contained at least one word from the list of the eleven most-frequent content words and at least one content word which occurred only once in the article. I was then in a position to compare the frequent words with the infrequent words in their tendency to be discarded in the syntactic pruning process.

Another thing which had to be decided in advance was some measure of probability of being discarded. Of course, one can have many different kinds of criteria for deciding which sentence elements to discard, and because of this one has to make a somewhat arbitrary choice of a measure. I decided that rather than attempt to work in terms of probability of being discarded, I would simply draw a dependency tree like that in Figure 1 for each chosen sentence, and would observe whether the frequent or the infrequent words occur higher in the tree. In order to make the idea "higher in the tree" a more definite notion, I put a metric restriction on the way in which the dependency tree could be drawn. The restriction, which confines all content words to numbered horizontal levels, is illustrated in Figure 2. The subject, verb, and object, if any, are placed on the topmost level, being the elements least likely to be dropped in the syntactic pruning process. Then on the second level come modifiers, dependent clauses, and other subordinate elements. Then on the third level come modifiers of the modifiers, and of the elements of the dependent clauses, etc. One can continue to go downward in level indefinitely, although in the 127 sentences selected from the two articles, none went below the sixth level; since about a dozen six-level sentences were observed, a rather sharp cut-off point is implied, a phenomenon which has been noted in the field of mechanical translation (14).

One has to adopt rules for deciding whether a given dependent element should be dropped to the next lowest level or retained on the same level. Here again, one can't help but be arbitrary in some of the rules: for example, it is possible to quarrel with the rule employed in the bottom sentence in Figure 2, by which the word "clouds" is dropped to the fourth level. Some would argue that "clouds" should be retained on the third level because transformation from the passive to the active would make it the subject. The opposite viewpoint would contend that the writer of the sentence chose the passive because he wanted to focus attention on the idea "light is obscured." Since we don't know, however, what effect either choice will have on the outcome of the investigation, one is safe in being arbitrary so long as the decision is made in advance and adhered to consistently. If one is strongly concerned about whether he has made the right choice, he can always repeat the experiment using the alternative rule.

FIGURE 2

Level

In the automatic indexing article, 77 sentences were found to contain a high frequency word (i.e., one of the top 11) and a word which occurred only once in the article. Out of these, 19 sentences had the infrequent word occurring on a lower level than the frequent word, 46 sentences had the reverse situation, and 12 sentences had the words on an equal level.* In the radio astronomy article, 50 sentences were found containing both the requisite frequent and infrequent words. Out of these, 23 sentences had the infrequent word on a lower level than the frequent word, 19 sentences had the reverse situation, and 8 had the words on an equal level.

If one considers the ratio of the number of sentences containing the infrequent words on a lower level to the number of sentences having the frequent words on a lower level, it is evident that a great variation exists between these two articles, at least. The ratio for one of the articles is approximately three times as large as that for the other. This finding, though not well-grounded methodologically, should still be enough to throw a scare into anyone considering the elimination of the so-called "subordinate elements" of sentences as a means of condensation. In the automatic indexing article especially, syntactic pruning would actually work at cross purposes to the Luhn method, because the most-frequent words have a strong tendency to gravitate to the lower levels of the dependency tree.

Admittedly one cannot generalize from one or two articles (although this would be far superior to generalizing from the counterexample); it is not the purpose of this paper to prove that the phenomenon exposed here is widespread, but only to reemphasize the idea that if variations in author behaviors can thwart automatic procedures, one had better study the extent and the effects of such variations before committing himself to the development of those procedures.

* If more than one frequent word (or infrequent word) is present in the sentence, their level numbers are averaged.

References

- (1) Luhn, H. P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development, 1957, 1, 309-317.
- (2) Zipf, G. K. Human Behavior and the Principle of Least Effort. Cambridge, Massachusetts: Addison-Wesley Press, Inc., 1949.
- (3) Parker-Rhodes, A. F. Contributions to the Theory of Clumps: The Usefulness and Feasibility of the Theory. ML-138, March 1961, Cambridge Language Research Unit.
- (4) Swanson, D. R. Searching Natural Language Text by Computer. Science, 1960, 132 (3434), 1099-1104.
- (5) Bar-Hillel, Y. Some Tehoretical Aspects of the Mechanization of Literature Searching. Technical Report No. 3. April 1960, U. S. Office of Naval Research, Washington, D. C.
- (6) Olney, J. FEAT, An Inventory Program for Information Retrieval. FN-4018, July 1960, System Development Corporation, Santa Monica, California.
- (7) Doyle, L. B. Indexing and Abstracting by Association. American Documentation, 1962, 13 (4), 378-390.
- (8) Katter, R. V. A Predictor of Semantic Communication Effect. TM-663/000/00, August 1962, System Development Corporation, Santa Monica, California.
- (9) Miller, G. A. and Selfridge, J. A. Verbal Context and the Recall of Meaningful Material. In Psycholinguistics, A Book of Readings edited by S. Saporta, New York: Holt, Rinehart and Winston, 1961, 198-206.
- (10) Hays, D. G. Automatic Language-Data Processing. In Computer Applications in the Behavioral Sciences, edited by H. Borko, Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1962, 394-421.
- (11) Simmons, R. F. Proto-Synthex Indexing System (Program Description). Unpublished memo, 1963, System Development Corporation, Santa Monica, California.
- (12) Doyle, L. B. Semantic Road Maps for Literature Searchers. Journal of the Association for Computing Machinery, 1961, 8 (4), 553-578.
- (13) Climenson, W. D., Hardwick, N. H., and Jacobson, S. N. Automatic Syntax Analysis in Machine Indexing and Abstracting. American Documentation, 1961, 12 (3), 178-183.
- (14) Yngve, V. H. Computer Programs for Translation. Scientific American, 1962, 206 (6), 68-76.

UNCLASSIFIED

System Development Corporation,
Santa Monica, California

THE MICROSTATISTICS OF TEXT.

Scientific rept., SP-1083, by
L. P. Doyle. 21 February 1963,
36p., 14 refs., 1 table.

Unclassified report

DESCRIPTORS: Documentation.
Data Processing Systems. Automation.

Presents a reappraisal of the role
of statistics in text analysis.

Discusses current inhibiting
influences in the use of statistics.

UNCLASSIFIED

Explores the question of descriptive
vs. predictive statistics. Makes a
distinction between macrostatistics
and microstatistics, with the
implication that the former should
be used in describing libraries whereas
the latter should be used in describing
written language. Pictures a relationship
between the probability of occurrence of
a word or word group in text and the cognitive
effect of such a word or word group.

Discusses the problem of selecting and
displaying content-indicative word groups in
condensed representations of documents.

Explains why the statistical approach, by
itself or in conjunction with other techniques
is unavoidable in a problem such as automatic
abstracting, and illustrates the perils faced
by some non-statistical methods which have been
talked about in recent literature.

UNCLASSIFIED